中国科学: 信息科学 2022 年 第 52 卷 第 11 期: 2075–2092
SCIENTIA SINICA Informationis
论文

附加偏见预测器辅助的均衡化场景图生成

王文彬[1,2], 王瑞平[1,2], 陈熙霖[1,2*]

1. 中国科学院计算技术研究所, 北京 100190
2. 中国科学院大学, 北京 100049
* 通信作者. E-mail: xlchen@ict.ac.cn

收稿日期: 2022–03–14; 修回日期: 2022–05–17; 接受日期: 2022–07–12; 网络出版日期: 2022–11–10

科技创新 2030 —“新一代人工智能” 重大项目 (批准号: 2021ZD0111901) 和国家自然科学基金 (批准号: U21B2025, U19B2036, 61922080) 资助项目

摘要 场景图是以场景中的物体为结点、以物体之间的关系为边构成的图结构, 在视觉与语言交互理解和推理相关任务中具有广泛的应用前景. 近年来, 场景图自动生成逐渐受到关注, 但生成结果中对于关系的描述受到长尾分布带来的偏见的影响, 偏向于样本量较大的头部关系. 然而头部关系往往过于空泛, 描述不够准确, 容易造成误解. 由于这种关系价值不高, 生成的场景图近似于退化为场景中物体信息的堆叠, 不利于其他应用在图结构上进行结构化推理. 为了使场景图生成器在这种不均衡的数据条件下, 能够更均衡地学习, 给出更加多样化的特别是尾部的更准确的关系, 本文提出一种附加偏见预测器 (additional biased predictor, ABP) 辅助的均衡化学习方法. 该方法利用一条有偏见的关系预测分支, 令场景图生成器抑制自身对头部关系的偏好, 并更加注重尾部关系的学习. 场景图生成器需要为指定的一对物体预测关系, 这是一种实例级的关系预测, 与之相比, 有偏分支以更简洁的方式预测出图像中的关系信息, 即不指定任何一对物体, 直接预测出图像中存在的关系, 这是一种区域级的关系预测. 为此, 本文利用已有的实例级的关系标注, 设计算法自动构造区域级的关系标注, 以此来训练该有偏分支, 使其具有区域级关系预测的能力. 在不同场景图生成器上应用 ABP 方法, 并在多个公开数据集 (Visual Genome, VRD 和 OpenImages 等) 上进行实验, 结果表明, ABP 方法具有通用性, 应用 ABP 方法训练得到的场景图生成器能够预测出更加多样化的、更准确的关系, 进而生成更有价值、更实用的场景图.

关键词 场景图生成, 长尾分布, 附加偏见预测器, 均衡化学习, 区域级关系

1 引言

真实场景中不仅包含物体, 也蕴含着物体之间丰富的关系, 而场景图就是一种以场景中物体为节点, 以物体间的关系为边的图表示. 场景图在计算机视觉领域中有着广泛的应用, 例如跨模态检

引用格式: 王文彬, 王瑞平, 陈熙霖. 附加偏见预测器辅助的均衡化场景图生成. 中国科学: 信息科学, 2022, 52: 2075–2092, doi: 10.1360/SSI-2022-0105
Wang W B, Wang R P, Chen X L. Balanced scene graph generation assisted by an additional biased predictor (in Chinese). Sci Sin Inform, 2022, 52: 2075–2092, doi: 10.1360/SSI-2022-0105

© 2022《中国科学》杂志社

图 1 (网络版彩图) Visual Genome [13] 数据库中的部分图片及相应的关系描述, 其中, 蓝色和红色关系词分别代表粗略和精细准确的关系词

Figure 1 (Color online) Some images from Visual Genome [13] and their corresponding relationship descriptions. The blue and red predicates stand for the coarse ones and precise ones, respectively

索 [1,2]、图像文本描述生成 [3,4]、视觉问答 [5,6] 和视觉推理 [7] 等. 因此, 关于场景图生成的研究日益受到关注 [6,8~12].

场景图的特点在于其能提供图像中各物体之间的关系, 这种关系描述越准确, 就越能够忠实地反映图像中物体之间的关联状态. 然而, 现有的场景图生成方法仍有很多不足. 这些方法识别出的关系大多平凡且图像特异程度不高, 比如现有方法 (如 Motif [10]) 对于图 1 [13] 中前两张图片都偏向于预测成 "on" 而非更准确的 "riding", "standing on" 等关系, 这虽然可以接受但不够准确, 很可能会产生歧义, 引发误解, 不利于场景图本身在其他应用中发挥作用. 究其原因, 就是现有的基于图像的场景图数据集中的关系标注普遍存在不均衡现象 [14,15], 频繁出现的关系特异性差, 而数量稀少的关系更加准确. 导致这种现象的原因主要有以下几点: (1) 不同关系词有数量不等的语义, 导致不同关系词的适用范围不同. 比如图 1 中所有样本用 "on" 描述都可以接受, 因为 "on" 是一种比较粗粒度的描述, 这使其具有更高的普适性和使用频率; 而更准确的关系由于描述更加细致、有针对性, 适用范围也更受限. (2) "报告偏差" (reporting bias) [16] 的存在 (即数据标注者在可接受范围内总是倾向于使用简单笼统的词), 导致本应可以用更准确的关系来描述的样本, 只要合理, 都被标注为更简洁但空泛的关系. (3) 更准确的关系词通常是一些动词, 而动词描述的是一个时间段内发生的动态事件, 对于静态的图像数据, 经常难以找到合适的动词来描述.

为了使模型在这种不均衡的数据条件下, 能够更均衡地学习, 预测出更加多样化的关系, 研究者们提出了多种类型的方法, 主要包括 (1) 重加权方法, 即给不同的样本施加不同的权重, 让模型着重学习样本数目少但更准确的关系 [17]; (2) 重采样方法, 即使用不同的采样策略来获取分布尽可能均衡的关系样本 [12]; 除此之外, 也有的方法在测试阶段对模型预测结果进行后处理调整 [14,15]. 这些方法都能够在一定程度上让模型预测出更多的尾部关系.

然而, 现有的大部分重加权方法在给不同样本施加不同权重时, 往往需要手动设计加权形式, 并调整较多的超参数; 而重采样方法往往使模型陷入对尾部关系的过拟合. 只在测试阶段进行后处理调整, 没有从本质上提升模型对于尾部关系的预测能力. 针对上述问题, 本文提出了一种附加偏见预测器 (additional biased predictor, ABP) 辅助的均衡化学习方法. 具体而言, 本文工作的目标是减轻场景图生成器在头部关系上的关注度, 进而更加注重学习尾部关系. 这种需求可以通过引入一条偏向于预测头部关系的有偏预测分支来实现. 对于头部关系, 有偏分支已经能够较好地将它们预测出来, 因此头部关系的损失对场景图生成器的影响降低, 场景图生成器无需再特别关注这些关系; 而对于尾部的关系, 有偏分支预测结果很差, 和没有该分支辅助的情况下相比, 场景图生成器需要更加注重对这些

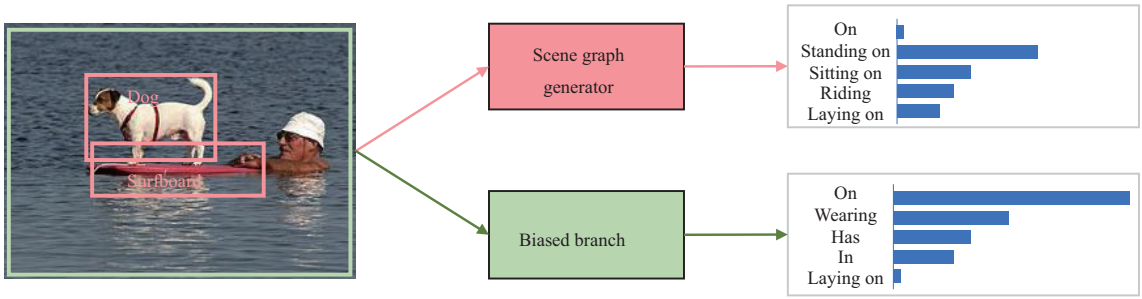

图 2 (网络版彩图) 场景图生成器与有偏预测分支各自完成的任务的示意图. 场景图生成器负责预测图片中某物体对之间的关系, 有偏分支负责在不指定物体对的情况下预测图片中可能存在的关系

Figure 2 (Color online) Tasks of the scene graph generator and the biased branch. The scene graph generator predicts relationships between a pair of objects, while the biased branch predicts possible relationships in the image without being assigned object pairs of interest

关系的学习, 尾部关系上的损失的影响增大. 本方法引入的有偏分支, 在训练过程中与场景图生成器形成 “互补”, 促使场景图生成器弥补在尾部关系上的不足, 进而更加均衡地学习. 和现有的重加权方法相比, 本方法以引入有偏分支的方式来抑制场景图生成器对头部关系的偏好. 而设计的分支简单直接且轻量级, 不包含超参数, 其中引入少量的可学习参数基本不影响场景图生成器的训练效率. 对于上述有偏见的预测分支, 最简单的设计方式就是以图像本身为输入, 输出该图可能存在的关系. 如图 2 所示, 不同于场景图生成器在指定物体 “dog” 和 “surfboard” 的前提下预测关系, 有偏分支在不指定物体对的情况下, 输出该图可能存在的 “on” (例如 “dog on surfboard”)、“wearing” (如 “man wearing hat”)、“in” (如 “man in sea”) 等关系. 此时, 因为没有指定要预测哪两个物体之间的关系, 这条分支只能依赖数据集中的 “图像 – 关系” 关联信息, 因此也会受到标注数据不均衡的影响. 由于这条分支需要完成的任务接近多标签分类, 本文利用已有的实例级的关系标注构建区域级的关系标注, 去训练这条轻量级的分支.

现有场景图生成器主要通过恰当地建模上下文信息来实现场景图的精确生成, 其中比较有代表性的三类建模架构包括 LSTM [10], TreeLSTM [6] 和 Transformer [18]. 本文在这几种类型的场景图生成器上应用 ABP 方法, 并在多个数据集 (Visual Genome [13], VRD [19], OpenImages [20] 等) 上开展实验. 结果表明, ABP 方法具有通用性, 能够有效地使上述不同生成器预测出更加多样化的关系, 进而获得更有价值的场景图.

本文的主要贡献有以下两方面. (1) 提出了通过添加偏见预测去辅助生成场景图的方法, 在原始场景图生成器的基础上, 实现了更好的类别均衡. 跨数据集情景下的实验表明, 本文提出的方法在准开放环境下仍然有效, 显示出其潜在的实用价值; (2) 通过对比实验, 考察了不同偏见项的影响, 验证了引入偏见项的有效性. 将本文方法与多种场景图生成器相结合, 仅增加少量的开销, 提升了原有场景图生成器的性能, 验证了其广泛的适用性, 且与重采样等均衡学习方法具有互补性和兼容性.

2 相关研究现状

在早期研究中, Johnson 等 [1] 提出使用场景图这种结构来进行跨模态检索, 由此研究者们开始注意到场景图的实用价值, 并研究自动生成场景图的方法. 稍早期的场景图生成研究集中在引入不同维度的信息并设计不同的信息传递机制来优化模型对于视觉关系的特征表示. 例如 Xu 等 [9] 设计了一种基于物体和关系之间的信息 “分发 – 汇合” 机制的场景图生成模型. Li 等 [21] 引入图像文本描述和

物体信息, 以多任务联合训练的形式来优化关系特征表示; Wang 等 [8] 提出可以利用相似的关系来增强模型的关系预测能力. 近期研究发现充分利用上下文信息是提高生成的场景图的质量的关键. 因此许多研究逐渐侧重于设计更好的上下文构建机制, 包括运用 LSTM 的 Motif [10] 和使用 TreeLSTM 的 VCTree [6] 和 HET [22]. Koner 等 [23] 使用了 Transformer [18], 而 Yang 等 [24] 和 Qi 等 [25] 使用了图神经网络.

上述方法在生成场景图时, 经常受到长尾分布引起的有偏预测问题 [14] 的困扰. 为了解决此问题, 可以参考在视觉识别领域使用的再均衡策略. 这些策略主要包括数据重采样, 使得不同类别的样本数目更加均衡 [26, 27], 以及根据实例的出现频率或者样本的难易程度对不同样本重加权 [17, 28]. 在物体检测和实例分割领域也有一些关于实例级的样本重采样方法 [29]. 然而一些工作 [30] 指出, 重采样方法改变了数据的分布, 可能会对模型的特征学习有害. 除此以外, 一些研究从迁移学习的角度出发, 将关于头部类别的判别特征迁移到尾部类别上 [31], 提升模型对尾部类别的识别能力. 基于这些通用策略, 一些研究工作做了专门针对场景图生成领域的适配. 例如, 根据某个结点的语义特性和出现频次来调整其损失权重 [32], 或者对样本进行图像级和实例级的两阶段采样 [12]. 还有的工作从场景图本身的来源考虑, 通过构建场景图和外部的常识或语言化知识的关联来增大尾部关系被预测的概率 [33, 34]. Suhail 等 [35] 提出了一个基于能量的模型训练框架, 基于不同关系之间的相互约束, 抑制被模型偏好预测的头部关系. Chiou 等 [36] 通过动态估计每种关系出现的频率, 从关系的有偏见的分布中恢复出不带偏见的分布. 利用关系之间非严格互斥的特性, Yu 等 [37] 构建树结构, 以由粗到细的方式指引模型将更多的注意力分配到尾部关系; 而 Guo 等 [15] 严格计算头部关系到尾部关系的转移概率并以之对预测出的头部关系进行调整, 使尾部关系有更多的机会被预测出来. 从因果分析的角度, Tang 等 [14] 发现单独依靠图像视觉信息产生的预测结果更加无偏.

不同于上述方法, 本文提出的方法不需要像重采样方法一样改变输入数据的分布, 也不像上述部分方法一样只在测试阶段对预测结果进行后处理, 而是在场景图生成器的训练阶段, 令引入的有偏预测分支与其形成 “互补”, 让场景图生成器更加注重对尾部关系的学习. 其中训练有偏分支的数据无需从外部引入, 而是利用现有标注自动构造.

3 方法

3.1 问题定义及方法背景

给定一张图像 $\mathcal{I}$, 场景图 $\mathcal{G}=\{\mathcal{O},\mathcal{R}\}$ 由图像中的物体集合 $\mathcal{O}=\{o_i=(c_i,\boldsymbol{b}_i)\}_{i=1}^{N}$ 和物体之间的关系集合 $\mathcal{R}$ 构成, 其中 $c_i\in\mathcal{C}$, $\mathcal{C}$ 表示所有物体类别的集合, $\boldsymbol{b}_i\in\mathbb{R}^4$ 表示物体边界框的坐标, 每条关系边 $r\in\mathcal{R}$ 具有标签 $p_{ij}\in\mathcal{P}$, p_{ij} 描述了物体 o_i 和 o_j 之间的关系, $\mathcal{P}$ 表示所有关系词的集合. 场景图生成即为对于给定图像 $\mathcal{I}$ 生成图 $\mathcal{G}$ 的过程.

本文所提出的 ABP 方法适用于目前主要的几类基于建模上下文信息的场景图生成器 [6, 10, 18]. 这些场景图生成器 $\mathcal{M}$ (在图 3 中以蓝色路径概括表示) 都包含以下部分.

特征图获取和物体位置提取. 为了获取图像中的物体信息, 预训练一个物体检测器, 例如, Faster R-CNN [38], 进而得到一组物体的边界框 $\{\boldsymbol{b}_i\in\mathbb{R}^4\}_{i=1}^{N}$ 和输入图像的 C 通道特征图 $\mathcal{F}\in\mathbb{R}^{C\times H\times W}$, 以及这组物体的初始类别标签 $\{l_i\}_{i=1}^{N}$, H 和 W 分别为特征图的高和宽.

物体信息编码与解码. 利用物体的边界框 $\boldsymbol{b}_i$ 在图像的特征图 $\mathcal{F}$ 上进行 RoIPooling [39] 操作获取每个物体的表观 (视觉) 特征 $\{\boldsymbol{v}_i\in\mathbb{R}^{d_v}\}_{i=1}^{N}$. 接下来经过一个物体特征编码器 $\mathrm{Enc}_{\mathrm{obj}}$ 融合每个物体的

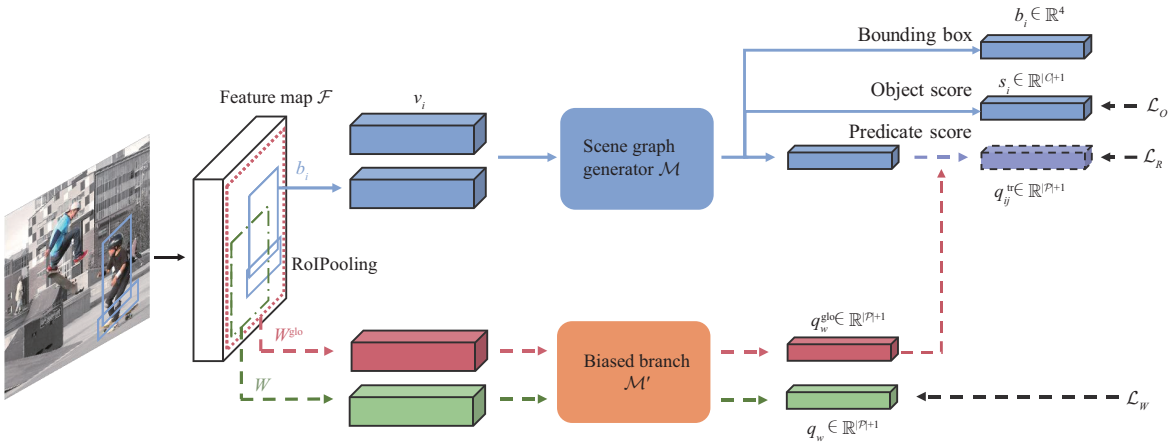

图 3 (网络版彩图) ABP 方法示意图. 训练阶段, 场景图生成器预测出的关系分数与有偏分支预测的关系分数 (红色路径) 相加; 绿色路径表示对有偏分支的训练过程. 推断阶段只保留场景图生成器预测出的关系分数. 虚线路径只存在于训练阶段

Figure 3 (Color online) Details of the ABP method. During the training stage, the predicate scores from the scene graph generator are added with those from the biased branch (the red path). The green path stands for the training process of biased branch. During inference, only those scores from the scene graph generator are retained. The dotted paths only exist in training stage

表观特征、位置特征和语义特征, 构建物体之间的上下文信息并通过信息传递机制获得改善后的物体特征 $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$:

$$\boldsymbol{x}_i = \text{Enc}_{\text{obj}}\left([\boldsymbol{v}_i, g(\boldsymbol{b}_i), \text{Emb}(l_i)]\right), \tag{1}$$

其中 $[\cdot]$ 表示将特征串接起来, Emb 表示可学习的特征嵌入层, 负责将类别标签转化为向量, $g(\cdot)$ 代表将边界框坐标映射成更高维度的特征向量的可学习函数. 这里的 Enc_{obj} 可以是 BiLSTM [10], BiTreeLSTM [6] 等. 然后经过解码器 Dec_{obj} (通常是一个由全连接层构成的分类器), 得到物体的各类别分数 $\boldsymbol{s}_i$ 和最终类别标签 c_i:

$$\boldsymbol{s}_i = \text{Dec}_{\text{obj}}(\boldsymbol{x}_i), \quad \boldsymbol{s}_i \in \mathbb{R}^{|\mathcal{C}|+1}, \tag{2a}$$

$$c_i = \arg\max_t \left[\boldsymbol{s}_i(0), \boldsymbol{s}_i(1), \ldots, \boldsymbol{s}_i(t), \ldots, \boldsymbol{s}_i(|\mathcal{C}|)\right], \tag{2b}$$

其中 $[\boldsymbol{s}_i(0), \boldsymbol{s}_i(1), \ldots, \boldsymbol{s}_i(t), \ldots, \boldsymbol{s}_i(|\mathcal{C}|)]$ 表示由 $\boldsymbol{s}_i$ 的元素构成的向量, $\boldsymbol{s}_i(0)$ 是其他的未知类的分数.

关系信息编码与解码. 使用一个与 Enc_{obj} 结构完全相同的编码器 Enc_{rel} 来从物体特征中获取应用于预测关系的特征 $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$:

$$\boldsymbol{y}_i = \text{Enc}_{\text{rel}}\left([\boldsymbol{v}_i, \boldsymbol{x}_i, \text{Emb}(c_i)]\right). \tag{3}$$

对于主体 o_i 和客体 o_j 之间的关系, 使用这两个物体的边界框形成的联合框 (最小矩形闭包) $\boldsymbol{b}_{ij}$ 在 $\mathcal{F}$ 上进行 RoIPooling 操作得到包含上下文信息的特征 $\boldsymbol{u}_{ij} \in \mathbb{R}^{d_u}$. 解码得到各个关系的分数 $\boldsymbol{q}_{ij}$ 和关系标签 p_{ij}:

$$\boldsymbol{q}_{ij} = \text{Dec}_{\text{rel}}(\boldsymbol{y}_i, \boldsymbol{y}_j, \boldsymbol{u}_{ij}) + (c_i \otimes c_j)\boldsymbol{W}^F, \quad \boldsymbol{q}_{ij} \in \mathbb{R}^{|\mathcal{P}|+1}, \tag{4a}$$

$$p_{ij} = \arg\max_t \left[\boldsymbol{q}_{ij}(0), \boldsymbol{q}_{ij}(1), \ldots, \boldsymbol{q}_{ij}(t), \ldots, \boldsymbol{q}_{ij}(|\mathcal{P}|)\right], \tag{4b}$$

其中 $[\boldsymbol{q}_{ij}(0), \boldsymbol{q}_{ij}(1), \ldots, \boldsymbol{q}_{ij}(t), \ldots, \boldsymbol{q}_{ij}(|\mathcal{P}|)]$ 表示由 $\boldsymbol{q}_{ij}$ 的元素构成的向量, $\boldsymbol{q}_{ij}(0)$ 是其他的未知关系的分数, $\mathrm{Dec}_{\mathrm{rel}}$ 代表关系解码器, $\boldsymbol{W}^F \in \mathbb{R}^{((|\mathcal{C}|+1)\times(|\mathcal{C}|+1))\times(|\mathcal{P}|+1)}$ 表示先验矩阵, 第 $r = c_i \times (|\mathcal{C}|+1) + c_j$ 行的向量 $\boldsymbol{W}_r^F$ 表示当主客体类别分别是 c_i 和 c_j 时, 关系的先验分布. $c_i \otimes c_j$ 产生一个长度为 $(|\mathcal{C}|+1)\times(|\mathcal{C}|+1)$ 的向量, 其中第 $c_i \times (|\mathcal{C}|+1) + c_j$ 个元素为 1, 其他元素为 0.

3.2 附加偏见预测器辅助的均衡化学习

上述场景图生成器 $\mathcal{M}$ 在学习过程中, 受到有偏数据的影响, 总是偏向于预测样本量更多的关系. 为了纠正这种偏见, 本文引入一条带有偏见的关系预测分支 $\mathcal{M}'$, 辅助 $\mathcal{M}$ 的训练过程. 此时 $\mathcal{M}$ 的关系预测分数变为

$$\boldsymbol{q}_{ij}^{\mathrm{tr}} = \boldsymbol{q}_{ij} + \boldsymbol{q}'_{ij}, \tag{5}$$

其中 $\boldsymbol{q}'_{ij}$ 表示 $\mathcal{M}'$ 的关系预测分数. 使用 softmax 交叉熵损失函数来训练 $\mathcal{M}$, 对于某个样本, 假设其关系标签为 y, 那么在加入分支 $\mathcal{M}'$ 前后单个样本的损失分别可以用下式中的 $\mathcal{L}_1$ 和 $\mathcal{L}_2$ 表示:

$$\mathcal{L}_1 = -\log \frac{\exp \boldsymbol{q}_{ij}(y)}{\sum_{k=0}^{|\mathcal{P}|} \exp \boldsymbol{q}_{ij}(k)}, \quad \mathcal{L}_2 = -\log \frac{\exp \boldsymbol{q}_{ij}^{\mathrm{tr}}(y)}{\sum_{k=0}^{|\mathcal{P}|} \exp \boldsymbol{q}_{ij}^{\mathrm{tr}}(k)}, \tag{6}$$

其中 $(\cdot)$ 表示取向量的某个元素. 对 $\mathcal{L}_1$ 和 $\mathcal{L}_2$ 作差可得

$$\begin{aligned} \mathcal{L}_2 - \mathcal{L}_1 &= \log \frac{\exp \boldsymbol{q}_{ij}(y) \sum_{k=0}^{|\mathcal{P}|} \exp \boldsymbol{q}_{ij}^{\mathrm{tr}}(k)}{\exp \boldsymbol{q}_{ij}^{\mathrm{tr}}(y) \sum_{k=0}^{|\mathcal{P}|} \exp \boldsymbol{q}_{ij}(k)} = \log \frac{\exp \boldsymbol{q}_{ij}(y) \sum_{k=0}^{|\mathcal{P}|} \exp \left(\boldsymbol{q}_{ij}(k) + \boldsymbol{q}'_{ij}(k)\right)}{\exp \left(\boldsymbol{q}_{ij}(y) + \boldsymbol{q}'_{ij}(y)\right) \sum_{k=0}^{|\mathcal{P}|} \exp \boldsymbol{q}_{ij}(k)} \\ &= \log \frac{\sum_{k=0}^{|\mathcal{P}|} \exp \left(\boldsymbol{q}_{ij}(k) + \boldsymbol{q}'_{ij}(k)\right)}{\exp \boldsymbol{q}'_{ij}(y) \sum_{k=0}^{|\mathcal{P}|} \exp \boldsymbol{q}_{ij}(k)} = \log \frac{\sum_{k=0}^{|\mathcal{P}|} \exp \left(\boldsymbol{q}_{ij}(k) + \boldsymbol{q}'_{ij}(k)\right)}{\sum_{k=0}^{|\mathcal{P}|} \exp \left(\boldsymbol{q}_{ij}(k) + \boldsymbol{q}'_{ij}(y)\right)}. \end{aligned} \tag{7}$$

当此样本的标签 y 属于头部关系时, 有偏分支 $\mathcal{M}'$ 预测出它为标签 y 的分数比其他绝大多数标签分数都大, 即 $\boldsymbol{q}'_{ij}(y) > \boldsymbol{q}'_{ij}(k)$, 此时 $\mathcal{L}_2 < \mathcal{L}_1$, 即在加入有偏分支 $\mathcal{M}'$ 后, 损失函数在头部关系样本上减小, 因此 $\mathcal{M}$ 无需再过度专注于这部分关系的学习; 相反, 当该样本属于尾部关系时, $\mathcal{L}_2 > \mathcal{L}_1$, $\mathcal{M}$ 需更加注重对这些关系的学习. 因此, 有偏分支 $\mathcal{M}'$ 与场景图生成器 $\mathcal{M}$ 形成 "互补", 使场景图生成器无需再把精力投入在头部关系上, 转而专注于学习那些有偏分支几乎预测不出来的尾部关系.

为了尽可能不影响场景图生成器的训练效率, 本方法令上述有偏分支完成更简单的关系预测任务, 即让其进行区域级的关系预测. 具体而言, 该分支基于一幅完整的输入图像直接预测图像中可能存在的关系. 因为不指定哪两个物体之间的关系, 分支很大程度上依赖标注中的 "图像 – 关系" 关联信息, 而关系标注本身就是不均衡的, 因此分支给出该图像可能存在的各种关系的分数也受到不均衡标注数据的影响. 而得到的关系分数随即被加到场景图生成器预测出的所有实例级的关系分数上.

根据上述思路, 本文提出的 ABP 方法具体如下: 对于图像 $\mathcal{I}$ 中的物体 o_i 和 o_j, 场景图生成器 $\mathcal{M}$ 根据式 (4a) 得到关系分数 $\boldsymbol{q}_{ij}$; 同时, 利用引入的有偏分支 $\mathcal{M}'$, 预测 $\mathcal{I}$ 中可能存在哪些关系, 得到相应的分数 $\boldsymbol{q}_w^{\mathrm{glo}} \in \mathbb{R}^{|\mathcal{P}|+1}$ (即为上述的 $\boldsymbol{q}'_{ij}$, 由于是图像级的分数, 不再加下标 ij 进行实例级的区分). 过程为: 使用一个和 $\mathcal{I}$ 尺寸相同的窗口 W^{glo} 在特征图 $\mathcal{F}$ 上提取特征并得到分数:

$$\boldsymbol{q}_w^{\mathrm{glo}} = \mathcal{Q}(\mathcal{F}, W^{\mathrm{glo}}) = \boldsymbol{W}_{\mathrm{cls}} \left(\boldsymbol{W}_2 \left(\boldsymbol{W}_1 \left(\mathrm{RoIPooling}\left(\mathcal{F}, W^{\mathrm{glo}}\right)\right)\right)\right), \tag{8}$$

其中 $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ 和 $\boldsymbol{W}_{\mathrm{cls}}$ 是可学习的参数. 而在 $\mathcal{M}$ 的训练阶段, 将 $\boldsymbol{q}_w^{\mathrm{glo}}$ 加到 $\boldsymbol{q}_{ij}$ 上, 得到最终预测结果 $\boldsymbol{q}_{ij}^{\mathrm{tr}}$ (如图 3 中的红色路径所示):

$$\boldsymbol{q}_{ij}^{\mathrm{tr}} = \boldsymbol{q}_{ij} + \boldsymbol{q}_w^{\mathrm{glo}} = \boldsymbol{q}_{ij} + \mathcal{Q}(\mathcal{F}, W^{\mathrm{glo}}). \tag{9}$$

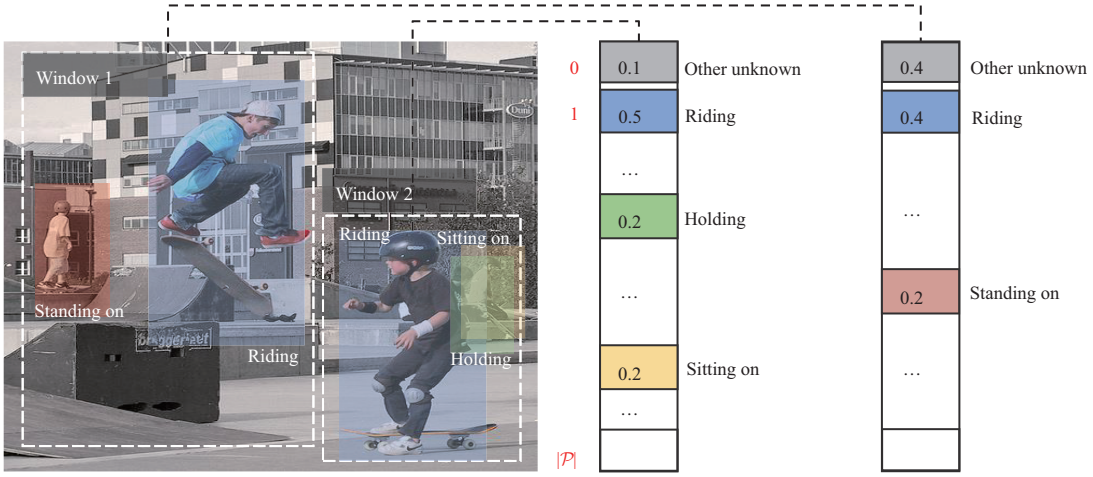

图 4 (网络版彩图) 窗口的关系软标签 (概率分布形式) 示意图, 其中每个窗口对应右侧的一个软标签, 标签内各种关系的概率值对应于关系实例区域 (用相似颜色表示) 与窗口的重合度

Figure 4 (Color online) Examples of the soft predicate labels (distribution) of windows. Each window has a soft label shown on the right side. The probability of each predicate in the soft label corresponds to the overlap between the relationship instance (shown with similar color) and the window

用 $\boldsymbol{q}_{ij}^{\mathrm{tr}}$ 来计算损失函数 (见 3.3 小节) 并训练 $\mathcal{M}$. 而在推断阶段, 关系预测结果只来源于场景图生成器 $\mathcal{M}$, 即最终推断结果为 $\boldsymbol{q}_{ij}^{\mathrm{inf}} = \boldsymbol{q}_{ij}$.

在实际实验中发现, 训练时如果将分支 $\mathcal{M}'$ 预测的一个批次内的图像的分数 $\boldsymbol{q}_w^{\mathrm{glo}}$ 随机打乱, 再与各个图像的关系实例分数 $\boldsymbol{q}_{ij}$ 相加, 能够增强场景图生成器的鲁棒性, 使其性能获得进一步提升.

有偏预测分支 $\mathcal{M}'$ 的训练及区域级关系标注构造. 由于分支 $\mathcal{M}'$ 以图像区域为输入, 输出图像内可能存在的多种关系, 类似于多标签分类过程, 受到 Lee 等 [40] 启发, 本文利用图像的实例级关系标注, 自动构建区域级的关系软标签. 具体而言, 为了提取更多的区域数据进行训练, 在图像上随机采样 N_w (实验中设定 $N_w = 8$) 个窗口, 每个窗口 W 的宽和高的最小值是 32 像素, 并且至少要涵盖一条标注关系. 然后, 对于窗口内涵盖的某种关系 t, 计算窗口与关系 t 包含的所有关系实例边界框的重合度, 作为窗口内出现关系 t 的概率, 详细计算过程如算法 1 所示. 最终获得的窗口的关系软标签 $\boldsymbol{l}^w$ 如图 4 所示. 最后, 使用采样的窗口 W 根据式 (8) 计算分数 $\boldsymbol{q}_w$, 并计算损失函数 (见 3.3 小节). 该损失仅仅用来训练分支 $\mathcal{M}'$ 的参数, 并不影响 $\mathcal{M}$ 的参数, 分支的训练过程如图 3 中的绿色路径所示.

3.3 损失函数设计

损失函数 $\mathcal{L}$ 包括训练场景图生成器 $\mathcal{M}$ 的损失 $\mathcal{L}_{\mathcal{M}}$ 和训练有偏分支 $\mathcal{M}'$ 的损失 $\mathcal{L}_{\mathcal{M}'}$ 两部分, 其中 $\mathcal{L}_{\mathcal{M}}$ 包括物体分类的 softmax 交叉熵损失和关系分类的 softmax 交叉熵损失, 分别计算如下:

$$\mathcal{L}_O = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{c}_i^{*\mathrm{T}} \log\left(\mathrm{softmax}\left(\boldsymbol{s}_i\right)\right), \quad \mathcal{L}_R = -\frac{1}{M}\sum_{k=1}^{M} \boldsymbol{p}_{ij}^{*\mathrm{T}} \log\left(\mathrm{softmax}\left(\boldsymbol{q}_{ij}^{\mathrm{tr}}\right)\right), \tag{10}$$

$$\mathcal{L}_{\mathcal{M}} = \mathcal{L}_O + \alpha\mathcal{L}_R, \tag{11}$$

$\boldsymbol{c}_i^*$ 是物体标签向量 (只有类别标签对应的那一维分量是 1, 其他分量全为 0), 共 M 条关系, 每条关系 r_k 的标签向量是 $\boldsymbol{p}_{ij}^*$.

算法 1 窗口的关系软标签构建

输入: 图像 $\mathcal{I}$, 标注的物体边界框 $\{\boldsymbol{b}_i\}_{i=1}^{N}$, 标注的关系 $\{r_k\}_{k=1}^{M}$, 窗口 W;
输出: 窗口的关系软标签 $\boldsymbol{l}^w \in \mathbb{R}^{|\mathcal{P}|+1}$;

$\boldsymbol{l}^w \leftarrow (|\mathcal{P}|+1)$ 维向量;
for $t \leftarrow 0$ to $|\mathcal{P}|$ **do**
  $\mathcal{S}_t \leftarrow \emptyset$;
  for $k = 1 \rightarrow M$ **do**
    $p_{ij} \leftarrow r_k$ 的类别标签;
    if $(t == 0)$ or $(p_{ij} == t$ and area$(\boldsymbol{b}_i \bigcap W) > 0$ and area$(\boldsymbol{b}_j \bigcap W) > 0)$ **then**
      $\mathcal{S}_t \leftarrow \mathcal{S}_t \bigcup \boldsymbol{b}_i \bigcup \boldsymbol{b}_j$;
    end if
  end for
  $\boldsymbol{l}^w[t] \leftarrow \sqrt{\mathbb{1}\{t == 0\}\text{area}(W) + (1 - 2\mathbb{1}\{t == 0\})\text{area}(W \bigcap \mathcal{S}_t)}$ $\quad \Leftarrow \mathbb{1}\{x\}$=1 当且仅当 x 为真;
end for
Return $\boldsymbol{l}^w/\text{sum}(\boldsymbol{l}^w)$.

$\mathcal{L}_{\mathcal{M}'}$ 为 softmax 交叉熵损失, 计算如下:

$$\mathcal{L}_{\mathcal{M}'} = -\frac{1}{N_w}\sum_{w=1}^{N_w} \boldsymbol{l}^{w\mathrm{T}} \log\left(\text{softmax}\left(\boldsymbol{q}_w\right)\right). \tag{12}$$

因此总体损失为

$$\mathcal{L} = \mathcal{L}_{\mathcal{M}} + \beta\mathcal{L}_{\mathcal{M}'} = \mathcal{L}_O + \alpha\mathcal{L}_R + \beta\mathcal{L}_{\mathcal{M}'}, \tag{13}$$

其中 α 和 β 是均衡因子, 将在 4.2 小节详细讨论并确定其取值.

4 实验结果

4.1 实验设置

数据集. 本文实验主要在以下常用的场景图数据集上进行. (1) Visual Genome [13]. 本文采用被使用较多的 Xu 等 [9] 处理的 VG150 版本. 该版本涵盖 150 类物体和 50 种关系, 训练集和测试集图片比例为 7:3, 按照惯例做法, 从训练集中分出 5000 张图片作为验证集. 同时, 为了展现在不同关系词上的性能, 本文采用 Li 等 [12] 的设定, 将关系词按照实例数目划分为 3 组, 即头部 (大于 10000 个实例)、中间 (500~10000 个实例) 和尾部 (少于 500 个实例). (2) VRD [19]. 数据集涵盖 100 类物体和 70 种关系, 共有 4000 张训练集图片和 1000 张测试集图片. (3) OpenImages [20]. 本文采用 Li 等 [12] 使用的 V6 版本, 数据集的训练集、验证集和测试集分别有 126368 张、1813 张和 5322 张图像, 涵盖 601 类物体和 30 种关系.

评测指标. 实验在 3 种标准的评测任务上进行 [9]: (1) 关系词分类 (predicate classification, PredCls), 即给定标注的物体框和类别, 只需要预测关系; (2) 场景图分类 (scene graph classification, SGCls), 即给定标注的物体框, 预测物体类别和关系; (3) 场景图生成 (scene graph generation, SGGen), 即需要同时检测物体和预测关系. 早期工作最常用的指标是 K 召回率 (R@K) [9, 10], 但该指标不能很好地反映模型在具有严重长尾分布特性的场景图数据集上的真实性能, 一般只要模型在头部类别上的性能好, R@K 的结果就会更好. 因此近期的工作主要报告并对比 K 平均召回率 (mR@K) [6, 12, 14, 15, 41], R@K

表 1 VG150 测试集上 ABP 方法在不同设置条件下的消融实验结果. 实验均以 Motif 为场景图生成器. 有 †标记的方法的实验结果由本文复现而来. 加粗和下划线分别代表最好和次好的性能

Table 1 Ablation study on VG150 with different configurations of ABP based on Motif. Methods with † mark are implemented by ourselves. The top-2 performances are shown with bold and underline, respectively

Experiment No.	Settings		PredCls (%)	
	Training	Inference	mR@20 / 50 / 100	R@20 / 50 / 100
1 (Motif [14])	✗	✗	11.5 / 14.6 / 15.8	59.5 / 66.0 / 67.9
2 (Motif†)	✗	✗	12.5 / 15.9 / 17.2	59.1 / 65.5 / 67.3
3 (Motif+TDE [14])	✗	Post-processing	18.5 / 25.5 / 29.1	33.6 / 46.2 / 51.4
4	Random	✗	12.5 / 15.8 / 17.2	59.4 / 65.7 / 67.5
5	Frequency	✗	20.2 / 27.1 / 30.0	39.2 / 50.3 / 53.5
6	Window	✗	<u>26.8</u> / <u>33.3</u> / <u>35.5</u>	35.4 / 42.3 / 44.1
7	Shuffled window	✗	**29.4** / **35.6** / **37.6**	34.2 / 40.3 / 41.9
8	Random	Window	8.4 / 10.5 / 11.3	57.7 / 63.9 / 65.6
9	Frequency	Window	12.6 / 15.9 / 17.2	56.5 / 63.6 / 65.5
10	Window	Window	<u>13.0</u> / <u>16.3</u> / <u>17.6</u>	58.1 / 64.7 / 66.4
11	Shuffled window	Window	**16.2** / **19.7** / **21.0**	56.8 / 63.3 / 65.1

不作为对比指标. 另外还采用了有图限制 (with graph constraint, 每对物体只允许预测一种关系) 和无图限制 (no graph constraint, 每对物体允许预测多种关系) [10] 两种准则.

实现细节. 实验中的超参数和 Tang 等 [14] 的设置一致. 受计算能力的限制, 处理的图像长度不大于 1000 像素, 否则将其归一化到 1000 像素. 在长度满足上述约束的前提下, 将宽度尽可能调整至不小于 600 像素. 上述归一化过程均保持比例不变. 采用以 ResNext-101-FPN [42] 为骨干网络的 Faster R-CNN [38] 作为前端检测器. 其训练独立于后续场景图生成部分的训练. 在训练场景图生成器和有偏预测分支时, 采用 SGD 作为优化器, 初始学习率为 1×10^{-2}, 每批次包含 12 张图像, 最大迭代次数为 50000 次. 当连续两次在验证集上的测试性能都没有上升时, 学习率衰减为原来的 1/10, 学习率的最大衰减次数为两次. 当学习率衰减次数超过最大衰减次数但尚未达到最大迭代次数时, 训练过程提前结束.

4.2 消融实验

以 Motif [10] 为场景图生成器 $\mathcal{M}$, 在 VG150 数据集上开展消融实验. 通过引入不同的附加偏见, 分析 ABP 方法的有效性和不同附加偏见的有效性. 具体参与对比的不同偏见项为: (1) 服从均匀分布的随机值 (Random), 即完全没有偏见 (对应表 1 中的实验 4); (2) VG150 上统计的各关系出现频率的对数值 (Frequency), 由于各关系的统计分布具有长尾特性, 因此天然地具有偏见 (对应表 1 中的实验 5); (3) 不打乱批次内的 $\boldsymbol{q}_w^{\text{glo}}$, 各自与对应的图像内的 $\boldsymbol{q}_{ij}$ 相加 (Window, 对应表 1 中的实验 6); (4) 打乱批次内的 $\boldsymbol{q}_w^{\text{glo}}$ 再与批次内各个图像内的 $\boldsymbol{q}_{ij}$ 相加 (Shuffled window, 即 ABP 的实现方式, 对应表 1 中的实验 7). 以上 4 项实验都遵循了 ABP 方法的设置, 即只在训练阶段引入有偏分支, 而最终预测结果不依赖有偏分支. 除此之外, 作为额外的观察, 也设置一组实验, 在推断阶段保留有偏分支的预测结果并评测此时的性能, 对应表 1 中的实验 8~11. 另外, 对于基线对比方法的选取, 本小节首先选择了 Motif 作为其中一个基线; 而 Motif+TDE [14] 方法比较典型, 和 ABP 方法正好相反, 不影响生成器的训练, 只在推断阶段对其预测结果进行后处理, 因此采纳为另一个基线对比方法.

ABP 方法的有效性. 将实验 5~7 等 3 项引入了偏见项的实验与基线 2 对比, 可以发现 ABP 方法在训练阶段引入有偏分支的做法是非常有效的. 使用不同的有偏的关系分数, 都使得 mR@K 有显著提升. 与之伴随的是 R@K 的一定程度的下降, 这是因为有相当一部分的头部关系被预测为含义相

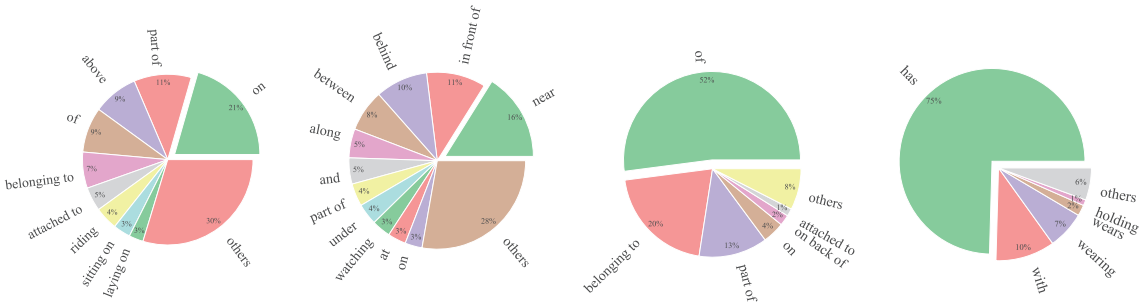

图 5　(网络版彩图) 部分标注的头部关系 (绿色突出部分) 被 ABP 方法预测为其他部分更准确的关系及其所占比例. 由于空间限制, 前两个图中占比低于 3% 的关系被归入 “其他” 部分

Figure 5　(Color online) The more precise predicates and their corresponding ratios predicted by ABP, which are originally annotated as the vague head predicates (the highlighted green parts). The predicates with ratios lower than 3% in the first and second charts are shown as “others” because of the restricted space

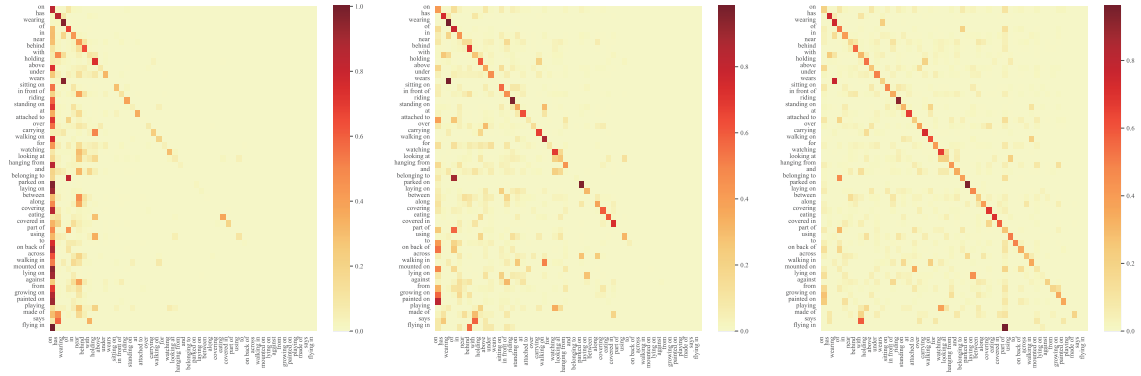

图 6　(网络版彩图) (a) Motif [10] 方法, (b) TDE [14] 方法和 (c) ABP 方法预测的关系的混淆矩阵

Figure 6　(Color online) The confusion matrices of relationship prediction from (a) Motif [10], (b) TDE [14] and (c) ABP

近的更准确的尾部关系, 而 R@K 指标显著地受头部关系的影响. 如图 5 所示, 位于头部的关系词 “on” 因为拥有不同的含义, 而被 ABP 合理地预测为许多意义相近的位于中后部的词, 比如表示 “从属” 含义时, 预测为 “attached to”, “part of” 等, 表示 “上面” 含义时, 预测为更准确的 “riding”, “standing on” 等. 更多的尾部关系被预测出来, 但 R@K 反而会下降, 这也正是 R@K 被 mR@K 取代为主要指标的原因. 这种现象在 TDE, BGNN [12], BA-SGG [15] 等现有方法中均能观察到. 与上述偏见项形成对比的是, 实验 4 中引入的均匀随机分布并不能发挥任何作用, 这也表明 ABP 方法有效的原因不在于引入了附加项, 而在于引入的附加项是有偏见的, 只有有偏见的附加项才能辅助场景图生成器更均衡地学习. TDE 方法与 ABP 不同, 其不影响生成器的训练过程, 只在其推断阶段进行一些后处理. 而 ABP 方法在训练阶段就让生成器更均衡地学习, 与只在推断阶段做结果后处理的 TDE 方法相比, ABP 明显对尾部的关系有更强的学习能力. 从图 6 中的混淆矩阵可以发现, 使用 ABP 方法的预测结果比基线 Motif 和 TDE 方法更加均衡.

不同偏见项的有效性. 对比实验 5~7 引入的 3 种有效的偏见项, 实验 5 中引入的是数据集级别的关系词的频率, 对于每张图像而言都相同, 性能最差; 而实验 6 和 7 引入区域级别的关系预测分数, 因此获得更好的性能; 其中实验 7 引入的被打乱的窗口分数使得有偏分支带来的预测结果不仅有偏,

表 2 PredCls 任务下使用不同 β 取值时的实验结果 (%), 以 Transformer 为场景图生成器并应用 ABP 方法

Table 2 PredCls results (%) based on the ABP-assisted Transformer with different settings of β

β	mR@20	mR@50	mR@100
0.1	31.2	38.6	41.2
0.2	31.1	38.3	40.9
0.5	30.7	37.7	40.5
1.0	31.2	38.2	40.1
2.0	31.5	38.1	40.5
5.0	30.9	38.3	40.7
10.0	29.7	36.4	38.8
20.0	29.3	35.9	38.5
50.0	26.3	32.4	34.9

而且和图像信息不对应, 这种更差的偏见项的引入增强了场景图生成器的鲁棒性, 使其进一步获得了约 2% 的性能提升. 最后, 作为额外的观察, 在推断阶段保留有偏分支时, 对比实验 8∼11 和基线 2, 发现 ABP 引入的基于区域的有偏预测分支 (实验 10 和 11) 也能轻微提升 mR@K, 并且 R@K 几乎保持, 这更多地归因于不同预测结果的叠加导致的性能提升.

均衡因子取值. 在式 (13) 定义的总体损失函数中存在两个均衡因子. 其中, 均衡因子 α 用于平衡损失 $\mathcal{L}_{\mathcal{M}}$ 中的 $\mathcal{L}_O$ 和 $\mathcal{L}_R$ 两种损失的影响. 在不同数据集上的实验 (见 4.3 小节) 表明将 α 设置为 1 可以获得稳定和良好的性能, 这也与现有大部分工作 [6, 10] 的设置一致. 分析其原因, 是因为在关注二元关系和物体时, 其数量级水平是相当的, 因此两项的权重也是差不多的. 本文将 α 设置为 1. 对于均衡因子 β, 以 Transformer [18] 为场景图生成器 $\mathcal{M}$, 在 VG150 数据集上考察其取值对性能的影响. 如表 2 所示. 结果表明, 当 β 取值在 0.1∼5.0 之间变动时, 对性能并没有显著的影响; 当 β 取值为 10.0 或以上时, 性能有较为明显的下降. 这是因为 β 取值较大时, 本质上场景图生成器 $\mathcal{M}$ 的作用消失, 对长尾的处理能力也相应丧失, 导致性能下降. 因此, 为了方便后续实验验证, 本文将 β 取值设置为 1, 也即假定有偏预测分支 $\mathcal{M}'$ 和场景图生成器 $\mathcal{M}$ 的贡献相当.

4.3 与既有方法的对比

ABP 方法适用于不同的场景图生成器. 现有的主流场景图生成器包括 Motif [10] 和 VCTree [6], 以及将 Transformer [18] 应用到场景图生成领域的一类生成器 [14, 23]. 本小节在这几种场景图生成器上验证 ABP 方法的有效性. 实验发现 ABP 搭配一些均衡采样方法可以进一步提升性能, 因此实验中也采用了两阶段重采样 (bi-level resampling, Birsmp) [12] 方法. 参与对比的既有工作包括 Motif [10], VCTree [6], GPS-Net [32], GB-Net-β [33], PCPL [43] 和 BGNN [12] 等. 除此以外, 还有大量的专注于均衡化预测的工作, 包括 TDE [14] 和 BA-SGG [15] 等. 在 VG150 上的实验结果如表 3 和 4 [44] 所示. 从结果中可以发现: (1) ABP 方法在 mR@K 指标上显著超越了既有方法. (2) ABP 和 Reweight, Birsmp 等传统的重加权、重采样方法相比, 更具优越性; 同时, ABP 也可以很好地和均衡采样方法配合使用, 能够获得更好的性能; 值得注意的是, 在 SGGen 任务设定下, 均衡采样方法和 ABP 配合使用的优势并不如在其他两种设定下明显, 可能原因是在此设定下, 需要将检测得到的候选关系对和标注的关系对进行匹配来给候选关系对分配类别标签 (根据物体边界框重合程度来匹配), 而检测得到的物体边界框并不精确, 即使重采样后有了更多尾部关系样例, 也可能没有匹配上候选关系对. 另外, 在 VRD 和 OpenImages 上的实验结果如表 5 和 6 所示 (其中 OpenImages 上其他方法的实验结果引自 Li 等 [12] 的工作), 这也体现了 ABP 方法具有广泛适用性.

表 3 VG150 测试集上的有图限制准则下的实验结果 (%). 第一组方法 (以横线分组) 以 VGG16 [44] 为骨架网络, 其他方法以 ResNeXt-101-FPN [42] 为骨架网络. 有 †标记的方法的实验结果由本文复现而来. 加粗和下划线分别代表最好和次好性能

Table 3 Results (%) of mR@K (with graph constraint) on the test set of VG150. Models in the first group (grouped by the horizontal lines) use VGG16 [44] backbone, while others use ResNeXt-101-FPN [42] backbone. Methods with † mark are implemented by ourselves. The top-2 performances are shown with bold and underline, respectively

Method	PredCls			SgCls			SGGen		
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
GPS-Net [32]	17.4	21.3	22.8	10.0	11.8	12.6	6.9	8.7	9.8
GB-Net-β [33]	–	22.1	24.0	–	12.7	13.4	–	7.1	8.5
PCPL [43]	–	35.2	37.8	–	18.6	19.6	–	9.5	11.7
BGNN [12]	–	30.4	32.9	–	14.3	16.5	–	10.7	12.6
Motif [10, 14]	11.5	14.6	15.8	6.5	8.0	8.5	4.1	5.5	6.8
Motif+Reweight [10, 14]	16.0	20.0	21.9	8.4	10.1	10.9	6.5	8.4	9.8
Motif+TDE [14]	18.5	25.5	29.1	9.8	13.1	14.9	5.8	8.2	9.8
Motif+Birsmp [10, 12] †	19.7	24.2	26.1	11.7	14.2	15.1	6.9	9.5	11.2
Motif+BA-SGG [15]	24.8	29.7	31.7	14.0	16.5	17.5	**10.7**	13.5	15.6
Motif+ABP	29.4	35.6	37.6	15.9	19.2	20.2	10.2	13.7	16.5
Motif+Birsmp+ABP	**31.0**	**36.5**	**39.0**	**18.6**	**21.6**	**22.9**	10.6	**14.5**	**17.5**
VCTree [6, 14]	11.7	14.9	16.1	6.2	7.5	7.9	4.2	5.7	6.9
VCTree+TDE [6, 14]	18.4	25.4	28.7	8.9	12.2	14.0	6.9	9.3	11.1
VCTree+Birsmp [6, 12] †	20.4	25.1	26.8	14.3	17.0	18.0	8.2	11.3	13.1
VCTree+BA-SGG [6, 15]	26.2	30.6	32.6	17.2	20.1	21.2	**10.6**	13.5	15.7
VCTree+ABP	29.8	36.3	38.4	19.5	23.0	24.1	10.2	**13.7**	**16.2**
VCTree+Birsmp+ABP	**31.0**	**36.8**	**39.0**	**23.0**	**26.8**	**28.5**	9.0	12.1	14.7
Transformer [14, 18] †	13.5	17.0	18.5	8.1	10.0	10.6	6.0	8.2	9.7
Transformer+Birsmp [12, 18] †	20.3	25.5	27.7	13.2	16.1	17.4	7.6	10.4	12.4
Transformer+BA-SGG [15]	26.7	31.9	34.2	15.7	18.5	19.4	11.4	14.8	17.1
Transformer+ABP	31.2	38.2	40.1	18.3	21.9	23.4	**11.5**	**15.4**	**18.3**
Transformer+Birsmp+ABP	**33.8**	**39.5**	**41.9**	**20.0**	**24.3**	**25.5**	11.2	14.8	17.5

表 4 VG150 测试集上的无图限制准则下的实验结果 (%). 第一组方法 (以横线分组) 以 VGG16 [44] 为骨架网络, 其他方法以 ResNeXt-101-FPN [42] 为骨架网络. 有 † 标记的方法的实验结果由本文复现而来. 加粗和下划线分别代表最好和次好性能

Table 4 Results (%) of ng-mR@K (with no graph constraint) on the test set of VG150. Models in the first group (grouped by the horizontal lines) use VGG16 [44] backbone, while others use ResNeXt-101-FPN [42] backbone. Methods with † mark are implemented by ourselves. The top-2 performances are shown with bold and underline, respectively

Method	PredCls			SgCls			SGGen		
	ng-mR@20	ng-mR@50	ng-mR@100	ng-mR@20	ng-mR@50	ng-mR@100	ng-mR@20	ng-mR@50	ng-mR@100
GB-Net-β [33]	–	44.5	58.7	–	25.6	32.1	–	11.7	16.6
Motif [10, 36]	19.9	32.8	44.7	11.3	19.0	25.0	7.5	12.5	16.9
Motif+Reweight [10, 14]	20.5	33.5	44.4	12.6	19.1	24.3	8.0	12.9	16.8
Motif+TDE [14, 36]	18.7	29.0	38.2	10.7	16.1	21.1	7.4	11.2	14.9
Motif+Birsmp [10, 12] †	27.9	43.2	55.5	16.5	24.9	31.5	9.6	14.9	19.6
Motif+ABP	34.0	48.8	59.9	19.3	27.6	33.2	12.1	18.1	23.1
Motif+Birsmp+ABP	**36.2**	**50.5**	**61.8**	**21.7**	**29.5**	**35.0**	**12.2**	**18.3**	**23.2**
VCTree [6, 14, 36]	21.4	35.6	47.8	14.3	23.3	31.4	7.5	12.5	16.7
VCTree+TDE [6, 14, 36]	20.9	32.4	41.5	12.4	19.1	25.5	7.8	11.5	15.2
VCTree+Birsmp [6, 12] †	28.8	43.4	55.6	20.1	30.5	38.3	11.4	17.4	22.9
VCTree+ABP	34.7	50.1	61.0	23.6	33.0	39.9	**12.3**	**18.2**	**23.0**
VCTree+Birsmp+ABP	**35.9**	**51.4**	**62.6**	**26.4**	**36.2**	**44.4**	10.7	15.9	20.8
Transformer [14, 18] †	20.4	34.7	46.7	12.9	21.1	27.8	8.2	13.6	18.7
Transformer+Birsmp [12, 18] †	28.2	43.6	56.1	18.2	27.5	34.3	10.5	16.7	21.7
Transformer+ABP	34.9	49.6	60.7	21.1	29.3	35.1	**13.2**	**19.6**	**24.5**
Transformer+Birsmp+ABP	**38.1**	**52.2**	**62.6**	**23.0**	**32.2**	**38.2**	13.0	18.7	24.2

表 5 VRD 测试集上 PredCls 任务的实验结果 (%). 模型均以 ResNeXt-101-FPN [42] 为骨架网络. 有 † 标记的方法的实验结果由本文复现而来. 加粗代表最好的性能

Table 5 PredCls results (%) on the test set of VRD. All models use ResNeXt-101-FPN [42] backbone. Methods with † mark are implemented by ourselves. The best performances are shown with bold

Method	mR@20	mR@50	mR@100	ng-mR@20	ng-mR@50	ng-mR@100
Motif†	8.3	9.6	9.8	14.2	25.3	36.4
Motif+ABP	**18.7**	**20.7**	**21.1**	**23.9**	**33.3**	**42.4**
VCTree†	9.2	10.3	10.6	16.5	29.2	42.2
VCTree+ABP	**20.1**	**21.8**	**22.2**	**25.6**	**37.2**	**49.0**
Transformer†	9.6	10.8	11.0	16.6	29.4	38.0
Transformer+ABP	**19.0**	**21.6**	**21.9**	**23.3**	**34.6**	**42.8**

表 6 OpenImages 测试集上 SGGen 任务的实验结果 (%). 模型均以 ResNeXt-101-FPN [42] 为骨架网络. 有 ∗ 标记的方法使用了重采样技术

Table 6 SGGen results (%) on the test set of OpenImages. All models use ResNeXt-101-FPN [42] backbone. Methods with ∗ mark apply the re-sampling strategy

	Motif	VCTree	G-RCNN [24]	GPS-Net [32]	Motif+TDE	BGNN [12]∗	Motif+ABP
mR@50	32.7	33.9	34.0	35.3	35.5	40.5	**39.8**

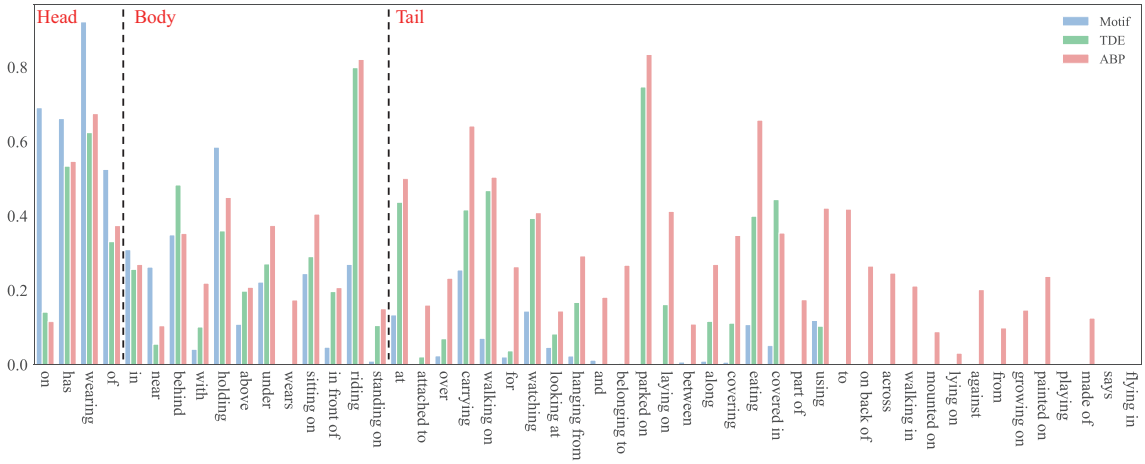

图 7 (网络版彩图) VG150 上不同方法的各关系词在 PredCls 任务下的 R@20 性能

Figure 7 (Color online) PredCls per-predicate R@20 results of different methods on VG150

为了更好地观察 ABP 如何影响每个关系的召回率, 图 7 中对比了不同方法在 VG150 上的每个关系的召回率. 相比基线 Motif, TDE 和 ABP 对不同类别的影响呈现相似的趋势, 都是在头部类别上下降, 在中部、尾部类别上提升. 但 ABP 对于尾部类别的性能提升更加显著, 尤其是在更靠近尾部的关系上, TDE 几乎没有任何作用, 而 ABP 仍然能够将它们正确预测出来.

图 8 展示了应用 ABP 后每个关系词的概率变化. 对于每个图片实例, Motif 预测的关系总是偏向于头部关系, 而正确的关系的概率因为没有头部关系的概率高而总被湮没. ABP 引入的有偏分支产生的概率同样偏向于头部关系. 在这个分支的辅助下, 场景图生成器将更多的注意力转移到尾部的关系学习上, 头部关系的概率被压制, 尾部关系的概率得以相对提升.

可视化结果. 图 9 展示了 Motif 和 Motif+ABP 方法产生的场景图实例. 其中有一些由 Motif 预测的关系是合理的, 比如第 1 个图中的 “giraffe near leaf”, 第 3 个图中的 “girl on skateboard” 和 “hair

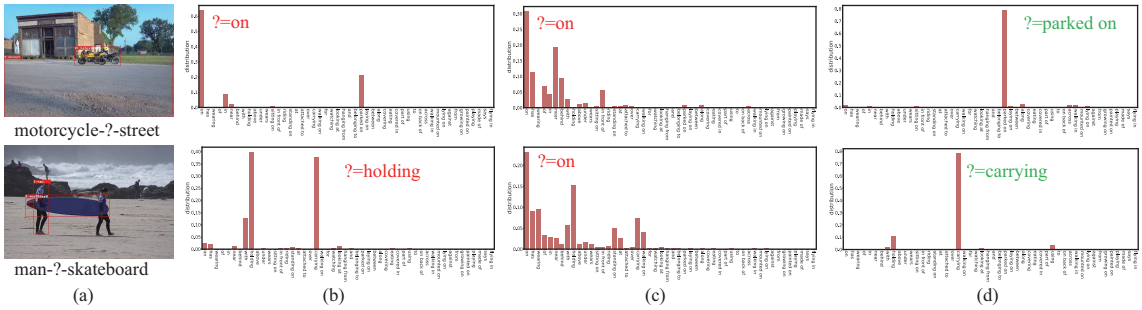

图 8 (网络版彩图) 从左至右分别为 (a) 关系样本以及 (b) 原始 Motif 生成器 $\mathcal{M}$、(c) 有偏预测分支 $\mathcal{M}'$、和 (d) $\mathcal{M}'$ 辅助训练的 Motif 生成器 $\mathcal{M}$ 产生的关系概率. 绿色和红色的关系词分别表示和标注一致或不一致

Figure 8 (Color online) (a) The samples and predicted probability distributions over the predicates from (b) original Motif $\mathcal{M}$, (c) the biased branch $\mathcal{M}'$, and (d) the Motif generator $\mathcal{M}$ trained with $\mathcal{M}'$. The green and red predicates are correct (GT) and incorrect (non-GT) ones, respectively

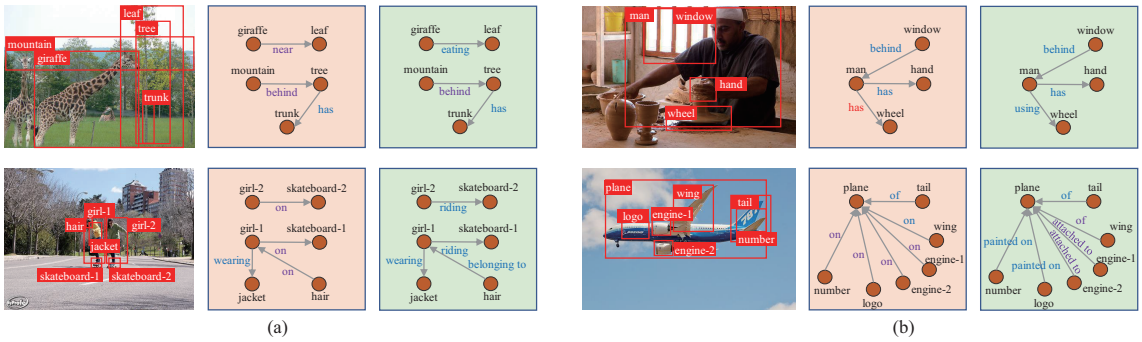

图 9 (网络版彩图) Motif (a) 和 Motif+ABP (b) 在 PredCls 任务设定下产生的场景图. 蓝色、紫色和红色的关系词分别表示该词与标注一致、与标注不一致但可以接受、与标注不一致且错误

Figure 9 (Color online) Scene graphs generated by Motif (a) and Motif+ABP (b) under the settings of PredCls. The blue, purple and red predicates are correct (GT), incorrect (non-GT) but reasonable, and incorrect (non-GT) and wrong ones, respectively

on girl", 第 4 个图中的 "number / logo / engine on plane", 但这些预测并没有对应上原本的标注; 而 Motif+ABP 将它们预测为 "giraffe eating leaf", "girl riding skateboard", "hair belonging to girl", "number / logo painted on plane" 和 "engine attached to plane", 这些关系词要么符合标注, 要么描述的准确程度更高 (实际上, 诸如 "number / logo / engine on plane" 等描述并不够精确, 因为 on 既可以根据语境被解释为 "附着在上面", 但也可以被理解成 "在 $\cdots$ 上面", 即使后面的情况更加罕见但也可能出现, 一旦出现的话就会造成误解). 另外, 一些关系被 Motif 预测错误, 比如第 2 个图中的 "man has wheel", 而 Motif+ABP 成功地将其预测为 "man using wheel".

时空开销. ABP 方法为场景图生成器引入有偏分支, 可能会产生额外的时间及空间开销. 表 7 中给出了不同模型的训练时间 (由于在推断阶段, 有偏分支被去除, 因此 ABP 方法不影响推断效率) 和参数量. 每个模型都在 2 块型号为 NVIDIA TITAN RTX 的 GPU 上进行训练, 每批次包含 12 张图像. 结合表 3, 4 和 7 等结果可以看出, ABP 方法在只增加少量的时空开销的情况下就能带来明显的性能增益, 具体表现为: 训练阶段单次迭代平均耗时只增加约 0.1~0.2 s, 参数量增加约 276 M (对于 Motif, VCTree 和 Transformer 而言, 分别相对增加 19.2%, 19.7% 和 21.4%), 在 VG150 上, Motif, VCTree 和 Transformer 等 3 种场景图生成器应用 ABP 方法后, 在 PredCls 任务下的 mR@20 指标获

表 7 PredCls 任务下模型的训练时间和参数量

Table 7 The training time under PredCls task and number of parameters

Method	Average time per iteration (s)	# Param (M)
Motif	0.79	1435
Motif+ABP	0.97	1711
VCTree	1.43	1401
VCTree+ABP	1.53	1677
Transformer	0.96	1292
Transformer+ABP	1.13	1568

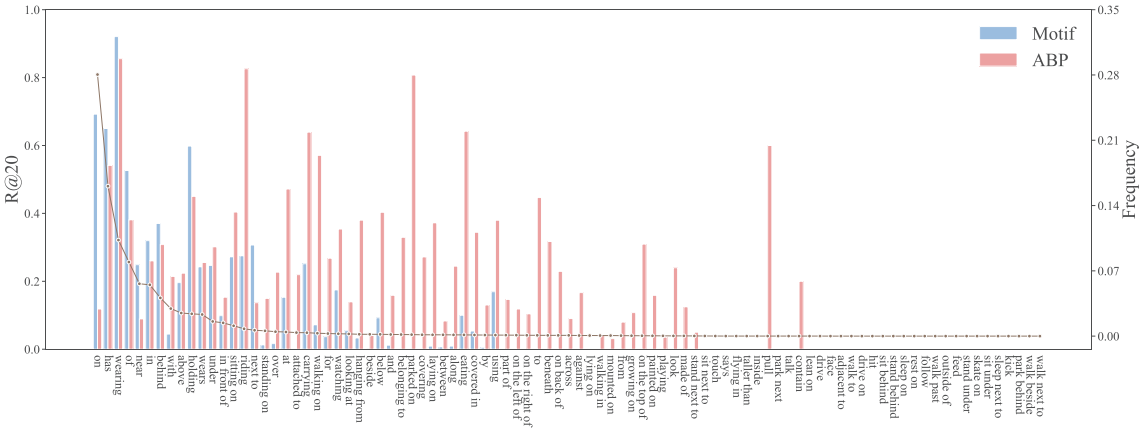

图 10 (网络版彩图) VGVRD 上不同方法的各类关系词在 PredCls 任务下的 R@20 性能. 折线表示各关系词的频率

Figure 10 (Color online) PredCls per-predicate R@20 results of different methods on VGVRD. The line represents the frequency of 91 predicates in VGVRD

得了 17%~18% 的绝对提升.

4.4 跨数据集验证实验

为了进一步验证 ABP 方法的均衡化效应, 需要验证其在跨数据集情景下的适用性, 但同时考虑到面向开放词表 [45] 的场景图生成研究仍处于起步阶段, 本小节以另一种方式进行验证. 具体而言, 将 VG150 和 VRD 的训练集与测试集对应合并, 对关系合并去重后得到 91 种不同关系. 合并后的数据集称为 VGVRD 数据集. 由于 VRD 中包含较多的更细粒度的关系词, 因此 VGVRD 受到更严重的长尾分布的影响, 如图 10 中折线所示. 首先, 在 VGVRD 上训练后测试, 结果如图 10 和表 8 中第 1 和 2 行结果所示. 虽然在部分非常接近尾部的关系词上使用 ABP 方法作用不大, 但和 Motif 相比, ABP 方法明显提升中后部的关系词的性能.

接下来进行跨数据集验证, 如表 8 所示, 对比表中第 3 和 4 行、第 5 和 6 行结果, 可以看到在跨数据集测试的情况下, 使用 ABP 方法仍然能够有效提升 mR@K, 验证了该方法的均衡化效应的广泛存在性.

5 结论与展望

针对场景图生成器受有偏数据影响而偏向于预测出现频率高但不准确的关系的问题, 本文提出了

表 8　PredCls 任务下的跨数据集实验结果 (%). 模型均以 ResNeXt-101-FPN [42] 为骨架网络

Table 8　PredCls results (%) of the cross-dataset experiment. All models use ResNeXt-101-FPN [42] backbone

Training set	Test set	Method	mR@20	mR@50	mR@100
VGVRD	VGVRD	Motif	8.0	10.0	10.7
		Motif+ABP	18.4	22.5	24.0
VGVRD	VG150	Motif	12.9	16.2	17.6
		Motif+ABP	27.8	33.8	36.1
VGVRD	VRD	Motif	10.7	12.1	12.4
		Motif+ABP	14.2	17.6	18.4

一种附加偏见预测器辅助的均衡化学习方法. 对于一个场景图生成器, 为其引入一条产生有偏结果的关系预测分支, 让这条分支辅助生成器进行学习. 由于在头部关系上, 有偏预测分支已经能给出较好的结果, 而在尾部关系上, 有偏预测分支表现很差, 这促使生成器更加注重学习出现频率低的尾部关系, 其预测结果也更加均衡. 通过消融实验, 验证了所提方法的有效性受益于偏见项的引入; 同时, 偏见项的引入仅在训练阶段增加少量的时间开销, 不影响推断效率, 而空间开销的增长与预测能力的提升相匹配. 将本方法与多种场景图生成器结合, 性能均有较明显的提升, 预测关系更加准确. 进一步, 在不同数据集上的实验以及跨数据集验证实验的结果表明, 本文提出的方法具有广泛的适用性.

场景图生成近年来得到了广泛关注. 虽然在标准评测数据集上, 方法的性能不断提升, 但仍有一些值得关注的问题有待进一步研究: (1) 生成场景图的最终目的是服务于后续的视觉理解应用, 当前大部分工作面向广泛的适用性要求, 直接使用场景图标注来训练模型, 常常与后续应用需求脱节, 需要考虑更好地平衡通用性与有效性; (2) 视觉关系描述具有不同的粒度, 需要进一步探索语义粒度可控的关系预测, 满足后续应用的不同需求.

参考文献

1 Johnson J, Krishna R, Stark M, et al. Image retrieval using scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 3668–3678

2 Wang S, Wang R, Yao Z, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2020. 1508–1517

3 Yao T, Pan Y, Li Y, et al. Exploring visual relationship for image captioning. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 711–727

4 Nguyen K, Tripathi S, Du B, et al. In defense of scene graphs for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 1407–1416

5 Antol S, Agrawal A, Lu J, et al. VQA: visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 2425–2433

6 Tang K, Zhang H, Wu B, et al. Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 6619–6628

7 Shi J, Zhang H, Li J. Explainable and explicit visual reasoning over scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 8376–8384

8 Wang W, Wang R, Shan S, et al. Exploring context and visual pattern of relationship for scene graph generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 8188–8197

9 Xu D, Zhu Y, Choy C B, et al. Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017. 5410–5419

10 Zellers R, Yatskar M, Thomson S, et al. Neural motifs: scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 5831–5840

11 Wang W, Wang R, Chen X. Topic scene graph generation by attention distillation from caption. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 15900–15910

12 Li R, Zhang S, Wan B, et al. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 11109–11119
13 Krishna R, Zhu Y, Groth O, et al. Visual Genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis, 2017, 123: 32–73
14 Tang K, Niu Y, Huang J, et al. Unbiased scene graph generation from biased training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 3716–3725
15 Guo Y, Gao L, Wang X, et al. From general to specific: informative scene graph generation via balance adjustment. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 16383–16392
16 Gordon J, van Durme B. Reporting bias and knowledge acquisition. In: Proceedings of the Workshop on Automated Knowledge Base Construction, San Francisco, 2013. 25–30
17 Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, Venice, 2017. 2980–2988
18 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 5998–6008
19 Lu C, Krishna R, Bernstein M, et al. Visual relationship detection with language priors. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 852–869
20 Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset V4. Int J Comput Vis, 2020, 128: 1956–1981
21 Li Y, Ouyang W, Zhou B, et al. Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE International Conference on Computer Vision, Venice, 2017. 1261–1270
22 Wang W, Wang R, Shan S, et al. Sketching image gist: human-mimetic hierarchical scene graph generation. In: Proceedings of European Conference on Computer Vision, 2020. 222–239
23 Koner R, Sinhamahapatra P, Tresp V. Relation transformer network. 2020. ArXiv:2004.06193
24 Yang J, Lu J, Lee S, et al. Graph R-CNN for scene graph generation. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 690–706
25 Qi M, Li W, Yang Z, et al. Attentive relational networks for mapping images to scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 3957–3966
26 Li Y, Vasconcelos N. REPAIR: removing representation bias by dataset resampling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 9572–9581
27 Li Y, Li Y, Vasconcelos N. RESOUND: towards action recognition without representation bias. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 513–528
28 Cui Y, Jia M, Lin T Y, et al. Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 9268–9277
29 Gupta A, Dollar P, Girshick R. LVIS: a dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 5356–5364
30 Zhou B, Cui Q, Wei X S, et al. BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 9719–9728
31 Liu Z, Miao Z, Zhan X, et al. Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 2537–2546
32 Lin X, Ding C, Zeng J, et al. GPS-Net: graph property sensing network for scene graph generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020. 3746–3755
33 Zareian A, Karaman S, Chang S F. Bridging knowledge graphs to generate scene graphs. In: Proceedings of European Conference on Computer Vision, 2020. 606–623
34 Zareian A, You H, Wang Z, et al. Learning visual commonsense for robust scene graph generation. In: Proceedings of European Conference on Computer Vision, 2020. 642–657
35 Suhail M, Mittal A, Siddiquie B, et al. Energy-based learning for scene graph generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 13936–13945
36 Chiou M J, Ding H, Yan H, et al. Recovering the unbiased scene graphs from the biased ones. In: Proceedings of the ACM International Conference on Multimedia, 2021. 1581–1590
37 Yu J, Chai Y, Wang Y, et al. CogTree: cognition tree loss for unbiased scene graph generation. 2020. ArXiv:2009.07526
38 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks.

In: Proceedings of Advances in Neural Information Processing Systems, Montreal, 2015. 91–99
39 Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1440–1448
40 Lee W, Na J, Kim G. Multi-task self-supervised object detection via recycling of bounding box annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 4984–4993
41 Chen T, Yu W, Chen R, et al. Knowledge-embedded routing network for scene graph generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 6163–6171
42 Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017. 5987–5995
43 Yan S, Shen C, Jin Z, et al. PCPL: predicate-correlation perception learning for unbiased scene graph generation. In: Proceedings of the ACM International Conference on Multimedia, 2020. 265–273
44 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
45 Zareian A, Rosa K D, Hu D H, et al. Open-vocabulary object detection using captions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 14393–14402

Balanced scene graph generation assisted by an additional biased predictor

Wenbin WANG[1,2], Ruiping WANG[1,2] & Xilin CHEN[1,2*]

1. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190*, China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049*, China*
* Corresponding author. E-mail: xlchen@ict.ac.cn

Abstract A scene graph is a structural representation of a scene comprising the objects as nodes and relationships between any two objects as edges. The scene graph is widely adopted in high-level vision language and reasoning applications. Therefore, scene graph generation has been a popular topic in recent years. However, it is limited by bias due to the long-tailed distribution among the relationships. Scene graph generators prefer to predict the head predicates, which are ambiguous and less precise. It makes the scene graph convey less information and degenerate into the stacking of objects, restricting other applications from reasoning on the graph. To make the generator predict more diverse relationships and provide a precise scene graph, we propose an additional biased predictor (ABP)-assisted balanced learning method. This method introduces an extra relationship prediction branch that is especially affected by the bias to make the generator pay more attention to the tail predicates rather than the head ones. Compared to the scene graph generator that predicts relationships between object pairs, the biased branch predicts the relationships without being assigned a certain object pair of interest, which is more concise. To train this biased branch, the region-level relationship annotation is constructed using the instance-level relationship annotation automatically. Extensive experiments on popular datasets, i.e., Visual Genome, VRD, and OpenImages, show that the ABP is effective on different scene graph generators. Besides, it makes the generator predict more diverse and accurate relationships and provides a more balanced and practical scene graph.

Keywords scene graph generation, long-tailed distribution, additional biased predictor, balanced learning, region-level relationship